UNIVERSIDAD DEL AZUAY

Facultad de ciencia y Tecnología

Escuela de Ingeniería en Alimentos

"Desarrollo de un método analítico para la cuantificación de lactosa en soluciones acuosas de varios azúcares"

Trabajo de graduación previo a la obtención del título de:

INGENIERA EN ALIMENTOS

Autora:

MÓNICA KATALINA GONZÁLEZ ZHICAY

Director:

Ing. Andrés Pérez González, Mgst.

CUENCA - ECUADOR

2022

DEDICATORIA

Esta tesis va dedicada a mi madre Luz Zhicay, por haberme dado la oportunidad de lograr culminar con una etapa más en mi vida, gracias por siempre apoyarme en todo y nunca dejarme sola, para mi es un ejemplo de lucha y superación constante.

A mis hermanos y sobrina de manera especial a Verónica por estar presente a lo largo de mi carrera Universitaria por ser el gran soporte que siempre necesité. A mi abuelita Leticia por ser mi segunda madre en la que siempre me refugié y por darme todo el amor incondicional durante toda mi vida.

Y a mi padre, que sé que desde el cielo festeja junto conmigo.

Sin todos ustedes no hubiese podido cumplir un sueño más, que hoy se hace realidad, espero siempre contar con todos ustedes.

Con mucho amor.

Mónica Katalina González Zhicay

AGRADECIMIENTOS

Primeramente, quiero agradecer a Dios por darme salud y vida y por permitir que hoy esté culminando con una etapa más.

De igual manera quiero expresar mi más sincero agradecimiento a mi tutor de tesis al Mgt. Andrés Pérez por toda su paciencia y dedicación y además por ser mi guía en el trayecto de mi tesis.

Al Dr. Piercosimo Tripaldi, la Dra. Diana Chalco y a todos los docentes que formaron parte de mi vida estudiantil, gracias por impartirme sus conocimientos con sabiduría y por formarme profesionalmente.

Finalmente agradezco a aquellas personas que siempre me apoyaron y que a pesar de todo estuvieron conmigo en las buenas y en las malas. Así mismo a mis compañeros de manera especial a Nicole y Heidy por siempre darme una mano amiga de manera incondicional y por hacer de mi periodo universitario único e inolvidable.

Desarrollo de un método analítico para la cuantificación de lactosa en soluciones acuosas de varios azúcares

RESUMEN

En esta investigación se desarrolló un método quimiométrico el cual permitió cuantificar la cantidad de lactosa presente en soluciones acuosas que contenían glucosa, fructosa y sacarosa por medio de la espectroscopia infrarroja (FTIR). El mejor modelo desarrollado para la lactosa fue mediante selección de variables con la aplicación de algoritmos genéticos con el método de regresión PLS el cual presentó valores de selección de R^2= 0.773 y Q^2= 0.726. El segundo mejor modelo fue mediante el uso del algoritmo (RSR) para selección de variables con el método de regresión KNN-5 presentando valores de selección de R^2= 0.802 y Q^2= 0.731.

Palabras claves: cuantificación de lactosa, lactosa, método analítico, espectroscopia infrarroja (FTIR).

Andrés Pérez González
DIRECTOR DE TESIS

María Fernanda Rosales
DIRECTORA DE ESCUELA

Mónica Katalina González Zhicay
AUTOR

Development of an analytical method for the quantification of lactose in aqueous solutions of various sugars

ABSTRACT

In this research, a chemometric method was developed. It made possible to quantify the amount of lactose present in aqueous solutions containing glucose, fructose and sucrose by means of infrared spectroscopy (FTIR). The best model developed for lactose was through selection of variables with the application of genetic algorithms with the PLS regression method, which presented selection values of R2= 0.773 and Q2= 0.726. The second best model was through the use of the algorithm (RSR) for the selection of variables with the KNN-5 regression method presenting selection values of R2= 0.802 and Q2= 0.731.

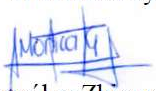
Keywords: lactose quantification, lactose, analytical method, infrared spectroscopy (FTIR).

Andrés Pérez González
THESIS DIRECTOR

María Fernanda Rosales
FACULTY DIRECTOR

Mónica Katalina González Zhicay
AUTHOR

Translated by



Mónica Katalina González Zhicay

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

La lactosa es uno de los disacáridos de gran importancia en los alimentos, este azúcar es el menos soluble y menos dulce; ya que, presenta el 15% del poder edulcorante de la sacarosa, varias personas a nivel mundial no lo toleran por la carencia de la enzima β-D-galactosidasa (lactasa); por lo que, a esta condición se la conoce como intolerancia a la lactosa. Su uso es muy frecuente en la industria alimentaria gracias a su poder adsorbente debido a que permite retener compuestos que generan sabores, aromas y colores (Badui, 2013)

Hoy en día, existen varias pruebas cualitativas y cuantitativas que se utilizan frecuentemente en los laboratorios para determinar azúcares reductores, las más utilizadas son: prueba de Benedict, reactivo de Fehling, reactivo de Tollens; mientras que a nivel industrial el método que más se aplica es el volumétrico de Lane-Eynon (Camacho, 2018). Sin embargo, estos métodos tradicionales de detección llegan a ser complejos, caros, toman mucho tiempo y no son específicos (Hernández et al., 2020). Por tales motivos, varios profesionales buscan diversas alternativas de análisis que sean rápidos, económicos y exactos, por lo que la espectroscopia infrarroja es una de ellas (Pérez, 2017).

Por su parte, el autor Pérez (Pérez, 2017) con su tesis titulada “Tratamientos matemáticos de señales espectrofotométricas infrarrojas para la cuantificación de azúcares” nos indica que en su trabajo, no se pudo obtener un buen modelo que permita determinar la cuantificación la lactosa debido a que este azúcar presenta una baja solubilidad con respecto a otros azúcares, además de compartir la misma zona del espectro por lo que sugirió que para encontrar un buen modelo sería necesario basarse en la solubilidad de este azúcar.

Por eso se ha planteado desarrollar un modelo analítico que permita cuantificar la lactosa mediante el cumplimiento de los siguientes objetivos.

- Realizar un diseño experimental de mezclas para evaluar la influencia que tiene cada azúcar sobre el espectro infrarrojo.

- Aplicar técnicas Quimiométricas para encontrar la mejor relación que existe entre los espectros infrarrojos y las concentraciones de las muestras del diseño experimental (regresión lineal).
- Desarrollar un modelo matemático que permita cuantificar la lactosa utilizando espectroscopia infrarroja (FTIR).

Para cumplir con dichos objetivos, se ha considerado utilizar 3 tipos de técnicas quimiométricas los cuales son: selección de variables, método de regresión y los valores de calidad del modelo R^2 y Q^2 los cuales permitirán validar los modelos para la lactosa, fructosa, glucosa y sacarosa.

1. CAPÍTULO 1: MATERIALES Y MÉTODOS

1.1. Espectroscopía Infrarroja (FTIR)

Esta metodología está representada por las siglas FTIR debido a que proviene del inglés, Fourier Transform Infra-Red o también puede llevar las siglas IR; esta técnica se fundamenta en la interacción de un haz de infrarrojo con la sustancia alimenticia a medir permitiendo obtener un interferograma, el cual mediante un algoritmo llamado transformada de Fourier se convierte en un espectro que va a estar representado por bandas y picos los cuales variarán en su forma de acuerdo a la naturaleza del alimento (Mondragón, 2017). Esta técnica es una excelente alternativa para los métodos analíticos debido a que el análisis es rápido, de fácil uso, confiable, la muestra necesita de poca preparación lo que conlleva a ahorros en el tiempo y en el costo, además de permitir analizar un mayor número de muestras (Rodriguez & Allendorf, 2011).

Por su parte, los autores Criollo y Cueva (Criollo & Cueva, 2017) detallan otros aspectos importantes los cuales se indican a continuación:

- La muestra a analizar debe ser muy pequeña.
- Se pueden analizar muestras sólidas, líquidas y gaseosas.
- El análisis es rápido, permitiendo obtener los espectros en poco tiempo.
- Es un método de análisis no destructivo.
- A comparación con los métodos tradicionales, la mano de obra es menor.

1.1.1. Métodos Analíticos FTIR

Los métodos analíticos FTIR se podrían dividir en tres tipos, en función de las zonas o regiones del espectro infrarrojo (IR):

1.1.1.1. Región Infrarrojo cercano (NIR)

Es un tipo de espectroscopia vibracional que se encuentra entre los 12500 a 400 cm^{-1}; en esta región la absorción se da cuando la frecuencia de vibración es igual a la frecuencia de radiación infrarroja dirigida a la molécula (Téllez, 2019). El NIR, es aplicable para cualquier molécula que tengan enlaces como: C-H, N-H, S-H, O-H (Criollo & Cueva, 2017). De igual manera, en esta zona se pueden observar que las bandas son anchas y tienen poca intensidad, lo que conlleva a dificultades visuales del analista conllevando a erróneas asignaciones de las bandas a estructuras o grupos funcionales; sin embargo, gracias el avance tecnológico que se ha dado con el pasar de los años ha permitido mejorar

la instrumentación específicamente en el desarrollo de softwares Quimiométricos los cuales han permitido tener una expansión considerable para los análisis de varias industrias (Valcárcel, 2009). Del mismo modo la autora Téllez, C (Téllez, 2019) expone que la posición de las bandas va a depender de ciertos factores como la temperatura, humedad, carácter cristalino o tamaño de las partículas de la muestra.

Esta técnica además de ser muy eficiente para análisis cuantitativos y cualitativos de productos alimenticios, no altera la muestra y permite obtener información de varios analitos con un solo espectro lo que le convierte en una técnica muy versátil (Flores, 2017). Así mismo, es un método limpio debido a que no se necesita de reactivos químicos ni tampoco se necesita seguir procedimientos analíticos para determinar varios compuestos (Valcárcel, 2009). En cuanto el ámbito laboral esta zona del espectro infrarrojo ha sido utilizada para el desarrollo de varias técnicas de análisis en la industria de los alimentos como es: el control de calidad de la leche, concentración de alcohol en vinos y bebidas de contenido alcohólico alto, muestras adulteradas con metanol o etilenglicol, entre otras (Maurad, 2016).

1.1.1.2. Región de infrarrojo medio (MIR)

Tiene como rango 4000-400 números de onda (cm^{-1}), esta región brinda información de la estructura molecular y los constituyentes de la muestra (Kuaquira & Huaman, 2022). En esta zona se pueden observar bandas correspondientes a cada estado vibracional de los enlaces y la intensidad de cada banda es proporcional a la concentración (ley de Lambert-Beer) por lo que es utilizada para análisis cuantitativo (Téllez, 2019).

Hoy en día la industria alimentaria frecuentemente hace uso de esta región, ya que sirve para diversos estudios como la medida de extractos fenólicos en vinos, control de originalidad y calidad de los quesos, caracterización de diversas etapas de maduración del queso camembert, determinación de ciertos parámetros químicos de los quesos emmental europeos como el nitrógeno total, nitrógeno soluble en agua y nitrógeno no procedente de proteína, determinación del origen botánico de mieles mediante la comparación de espectros, permite detectar adulteraciones en la carne cocida, entre otros (de Fuentes Navarta et al., 2008).

Anteriormente, su aplicación era limitada debido a las grandes absorciones de agua en todo el espectro dando como resultado la difícil preparación y manipulación de la muestra (Van de Voort, 1992), pero gracias al desarrollo de instrumentos que abarca la

transformada de Fourier (FTIR) se ha logrado mejorar la calidad de los espectros y al mismo tiempo, se ha reducido el tiempo para la obtención de datos (Serrano, 2017). De igual manera los autores Billabio y Todeschini (Billabio & Todeschini, 2009) hacen referencia que esta región se implementa como una manera de realizar mediciones no destructivas y no invasivas por lo que la preparación de la muestra y la recopilación espectral es mínima.

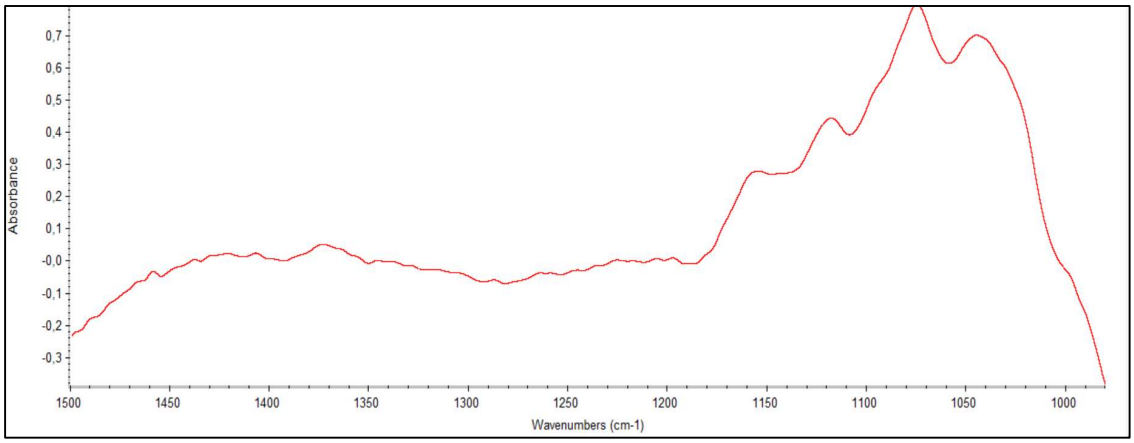


Figura 1. Espectro Infrarrojo de la Lactosa en un intervalo de 1500-980 cm^{-1}.

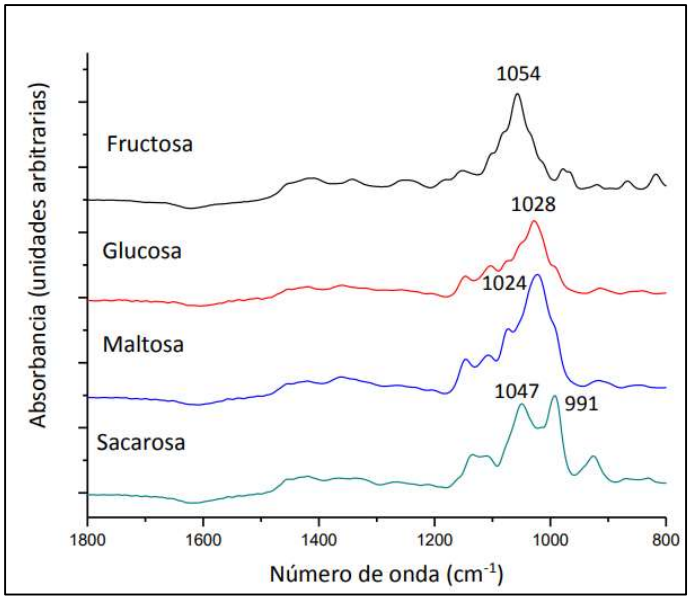


Figura 2. Espectros FTIR-ATR de diferentes azúcares disueltos en agua en intervalos de 1800-800 cm^{-1}.

1.1.1.3. Región de infrarrojo lejano (FIR)

Esta zona se encuentra entre los 400-10 números de onda (cm^{-1}), en esta sección se va a tener información conformacional de las estructuras (Pérez, 2017). En cuanto al uso, esta zona es utilizada para realizar análisis de compuestos orgánicos, inorgánicos u organometálicos que tengan átomos pesados cuya masa atómica sea mayor a 19 (Serrano, 2017).

1.1.2. Métodos tradicionales de análisis de azúcares

1.1.2.1. Método volumétrico general de Lane-Eynon

Se lo realiza de manera manual por lo que no es automatizado, consiste en analizar los azúcares reductores mediante titulación tomando como ventaja el poder reductor de los azúcares; este método se basa que en un medio alcalino el cual va a estar conjuntamente con la solución de Fehling se va a realizar la reacción del sulfato cúprico con el azúcar reductor llevando a la formación de oxido cuproso; el cual, posteriormente presentará un precipitado de color rojo ladrillo, en este método es indispensable utilizar el azul de metileno; ya que, servirá de indicador y este será decolorado cuando el cobre se haya reducido en su totalidad dándonos como indicativo que la titulación ha terminado (Contreras et al., 2019).

1.1.2.2. Método de Munson y Walker

Es un método gravimétrico utilizado para determinar la concentración de azúcares reductores, esta técnica consiste en hacer hervir por aproximadamente 4 minutos cantidades conocidas de reactivo de Fehling y la solución de azúcar en un matraz. A medida que transcurre este tiempo se va a dar la precipitación del oxido de cobre 1 (Cu_2O) llegando a obtener un sólido el cual se va a filtrar y se lavar en el siguiente orden (agua a 60 ºC, alcohol y éter), finalmente se va dar nuevamente la formación de un sólido el cual se va a secar a 110ºC hasta peso constante y partiendo del peso de Cu_2O se obtendrá la cantidad de azúcar (Camino et al., 2020).

1.1.2.3. Cromatografía líquida de alta resolución (HPLC)

Es un tipo de cromatografía en columna ampliamente utilizada para separar los componentes de una muestra por medio de distintas interacciones químicas entre la columna cromatográfica y la sustancia analizada (Sarria, 2018). En cuanto a los análisis de azúcares este tipo de cromatografía tiene diversas ventajas debido a la solubilidad en

el agua, sensibilidad y fácil adaptación a determinaciones cuantitativas exactas (Herrera, 2011). Sin embargo, puede presentar diversos problemas conduciendo a resultados erróneos como es el cambio de la composición de la fase móvil, cambios de flujo, fluctuaciones de la temperatura, falta de equilibrio entre los análisis sucesivos y la inyección del tamaño de las muestras (Zumbado, 2010).

1.1.2.4. Método de cobre de Roberts

Este método ha sido utilizado para realizar la hidrolisis enzimática del polisacárido dextrano conjuntamente con el método de Haze, debido a que este polisacárido no es deseable cuando se da el procesamiento del azúcar y siendo producido mediante la contaminación de ciertos microorganismos influyendo negativamente en el producto final. Básicamente el método de cobre de Roberts permite medir la concentración de dextranos más grandes que trisacáridos (Abraham et al., 2014).

1.1.2.5. Métodos colorimétricos

Son utilizados para la determinación de azúcares reductores, este análisis se basa en la reacción de estos compuestos con determinados reactivos dando como resultado los derivados coloreados. Se toma como resultado positivo cuando se da la formación de color indicando la presencia del compuesto, y resulta negativo cuando no se da ninguna formación de color (Úbeda, 2012).

1.1.2.6. Análisis enzimático

Comúnmente se utilizan dos tipos de análisis para la detección específica y cuantificación de azúcares, los cuales son: específico para monosacáridos y específico para la hidrólisis de oligosacáridos de cadena larga; sin embargo, el método es limitado debido a la presencia de contaminantes que se encuentran en las disoluciones de prueba o por la presencia de otros azúcares, sales y metales. De igual manera, este método es contraproducente; ya que, cada azúcar requiere de diferentes medios para el análisis, así como de varias enzimas (Úbeda, 2012).

2. CAPÍTULO 2: Diseño Experimental

Hoy en día, dentro del campo industrial es habitual hacer experimentaciones o pruebas que permitan resolver problemas o esclarecer una idea dada por otras entidades (hipótesis). Es por eso que el diseño experimental permite planear y realizar diversas pruebas y como estas deben efectuarse y de qué manera para poseer datos que al ser analizados estadísticamente brinden evidencias objetivas permitiendo responder las interrogantes que se plantearon en un inicio sobre una determinada situación con el fin de resolver el problema o mejorarlo (Gutiérrez & de la Vara, 2008).

Tiene una amplia aplicación en diferentes áreas disciplinarias como: las ingenierías, debido a que sirven para el control de la calidad, métodos de predicción y control de procesos, además de mejorar el rendimiento de un proceso de manufactura; en la Física en cuanto a la teoría cinética de los gases; en la psicología en cuanto a la personalidad, conducta, inteligencia; entre otras ramas más (Rojas & Rojas, 2000).

2.1. Diseño experimental de mezclas

Emplea el criterio que la suma de las proporciones de los componentes es el 100% y la modificación de uno de ellos va a afectar al resto, este diseño ha venido aplicándose con éxito en la industria alimentaria y farmacéutica debido a que permite evaluar la influencia de cada componente de la mezcla en el producto final, además su aplicación ha permitido optimizar las proporciones de diversas frutas en bebidas y en otros productos (Bayas & Saltos, 2010).

Se pueden utilizar dos tipos de diseño: el primero es el llamado diseño simplex -lattice el cual es aplicable cuando los datos se distribuyen sobre una región de superficie de respuesta y el segundo, es el diseño simplex-centroid el cual se usa cuando los datos se distribuyen alrededor del centro de la región de superficie de respuesta (Delgado, 2008).

Con el pasar del tiempo, el diseño experimental ha ido evolucionando, llegando a convertirse en una de las disciplinas con mayor importancia en cuanto a las investigaciones experimentales debido a que permite tener un mayor control en la variación de los datos. La experimentación tiene como objetivo principal analizar todas las fuentes que presenten variación en los datos, permitiendo la visualización de la dependencia de las variables dependientes e independientes de tal manera que permita establecer conclusiones válidas y objetivas (Reyes, 2009).

3. CAPÍTULO 3: Quimiometría

El análisis multivariado de datos (Quimiometría) son métodos matemáticos y estadísticos que sirven para brindar información de muestras partiendo de sus datos químicos. La aplicación de técnicas quimiométricas es de mucha utilidad en el estudio de espectros, en este caso infrarrojos, debido a la complicada interpretación de los espectros por la naturaleza de las bandas, la superposición de unas bandas con otras y al desplazamiento de bandas por la presencia de puentes de hidrógeno, que tienden a ser bandas débiles y anchas; es fundamental aplicar estos métodos de análisis ya que permiten obtener información útil ignorando el ruido (Flores, 2017).

En cuanto a las áreas de trabajo que aplica esta disciplina se encuentra el reconocimiento de patrones, calibración multivariada, estudios de QSAR, procesamiento de señales químicas, resolución matemática de mezclas complejas, diseño de experimentos y los análisis de imágenes químicas (Domínguez, 2014).

3.1. Métodos Quimiométricos

Para el presente estudio podemos considerar que se pueden clasificar en cuatro grupos de acuerdo a su uso:

3.1.1. Pretratamientos de espectros o pretratamientos matemáticos

Estos pretratamientos se aplican cuando los datos son de origen espectroscópico (Torres, 2016). El objetivo es buscar que los espectros que se obtuvieron, permitan realizar un correcto desarrollo de los modelos cualitativos y cuantitativos los cuales se pueden obtener mediante el aislamiento de señales favorables y desfavorables, eliminación del ruido instrumental, suavización de los datos, desplazamiento de la línea de base, linealización de los datos, entre otras (Flores, 2017).

A continuación, se detallan los pretratamientos más utilizados en el tratamiento de espectros:

- Suavización: Permite eliminar los picos angostos de los espectros para tener un menor ruido aleatorio.
- Corrección de línea de base: Existen dos maneras para realizar esta acción; la primera consiste en corregir el espectro mediante la aproximación de la línea base a una función polinómica la cual posteriormente se le restará al espectro; y la segunda consiste en realizar el cálculo de la primera y segunda derivada del

especto con la finalidad de no tener desplazamientos de la línea base (Torres, 2016).

- Autoescalado: Es un vector que normaliza las dimensiones de las muestras a media cero y desviación típica, los métodos de normalización permiten eludir variaciones experimentales y disminuyen errores computacionales cuando se emplean métodos de reconocimiento de patrón (Valcárcel, 2009).

3.1.2. Análisis exploratorio de datos o Análisis por componentes principales

Se emplea antes de desarrollar el modelo de regresión, el cual va a permitir encontrar similitudes entre las muestras, además de identificar a las que no estén dentro del grupo; en este análisis el resultado se visualiza como gráficos siendo el Análisis de Componentes Principales (PCA) la más utilizada (Flores, 2017). El PCA consta de varios objetivos cuando existe una correlación entre las variables, los cuales son (Torres, 2016):

- Permitir crear nuevas variables las cuales tendrán información y la expresarán como un conjunto de datos.
- Eliminar las variables que carezcan de información debido a que no servirá de aportación en el estudio que se esté realizando.
- Proporcionar mayor facilidad de interpretación de los datos.
- Disminuir la dimensionalidad de los datos (Torres, 2016).

3.1.3. Modelos para análisis cualitativos: discriminación y clasificación

Son técnicas utilizadas para la identificación de patrones, permite analizar la región de infrarrojo cercano (NIR) de una muestra conocida con una desconocida con la finalidad de encontrar semejanzas y diferencias. En el campo alimentario es útil para determinar una adulteración y el grado de pureza de mieles y aceites como el de oliva (Flores, 2017).

Existen diversas técnicas utilizadas para el desarrollo de modelos de clasificación, las más empleadas son: análisis de grupos, análisis discriminante PLS (PLS-DA) análisis de variación canónica (CVA), análisis de k-cercano, análisis discriminante lineal (LDA), análisis discriminante factorial (FDA), entre otros (Flores, 2017).

3.1.4. Modelos para análisis cuantitativos: regresión y predicción.

La región NIR al ser una técnica analítica secundaria, necesita de una correlación para poder medir una propiedad química de una muestra, esta correlación se da entre el

espectro y el valor de esa propiedad o también conocido como valor de referencia el cual se resuelve por medio de un método de análisis independiente.

Los métodos más utilizados son: regresión de componentes principales (PCR), mínimos cuadrados parciales (PLS) y la regresión múltiple lineal (MLR) teniendo en común el uso de técnicas de mínimos cuadrados (Flores, 2017).

3.2. Métodos de selección de variables

3.2.1. Algoritmos genéticos (GA)

Son modelos computacionales que permiten resolver problemas de optimización; además, de ser considerados métodos adaptivos los cuales se basan en la recombinación genética de los organismos vivos. Este método utiliza el concepto de población, en donde cada individuo es representado por una solución y a la misma vez es candidato para un problema de optimización dado. Las soluciones más prominentes van a poseer una mayor oportunidad de ser seleccionadas para reproducirse, a este proceso de reproducción se le llama cruzamiento; la recombinación permite crear nuevas soluciones las cuales poseen información genética de los padres e información genética inherente a la propia naturaleza de dicho individuo (Cuevas et al., 2021).

En cuanto al ámbito de aplicación, es utilizado para problemas de optimización numérica debido a la complicada resolución con los métodos tradicionales, estos algoritmos se subdividen en tres categorías encontrándose los problemas de secuenciación, problemas numéricos y problemas de selección de subconjuntos (Massart et al., 1998).

3.2.2. Algoritmo de remplazo secuencial renovado (RSR)

Cassotti, Grisoni y Todeschini desarrollaron el algoritmo RSR tomando como base el algoritmo de remplazo secuencial (SR) el cual remplaza cada una de las variables que se encuentran en un modelo de tamaño M (con $M<p$) con cada variable restante para obtener un mejor modelo, sin embargo, este modelo no realiza todas las combinaciones posibles de las p variables lo que conlleva a un menor tiempo de análisis y siendo menos metaheurístico (Cassotti et al., 2014). Estos autores le agregaron 4 funcionalidades las cuales se detallan a continuación:

a) Menor tiempo para realizar el cálculo.
b) Adición de herramientas de validación los cuales se refirieron a la capacidad predictiva de los modelos.

c) Mayor probabilidad de encontrar al mejor modelo.
d) Seleccionar modelos con patologías (Pérez, 2017).

3.3. Modelos multivariantes de regresión

3.3.1. Mínimos cuadrados ordinales (OLS)

Es el método más sencillo; ya que, este tipo de regresión relaciona los parámetros y predictores de manera lineal, por lo que está representado por la siguiente ecuación:

$$y = Xb + e$$

Para poder realizar los cálculos de los coeficientes de regresión **b,** es necesario resolver la siguiente ecuación (Pérez, 2017).

$$\hat{b} = (X^T X)^{-1} X^T y$$

Donde:

X= matriz de variables independientes

y= respuesta

e = constante

Los exponentes de la segunda ecuación hacen referencia al cálculo matricial.

T = Transpuesta

-1 = es la inversa (Pérez, 2017).

3.3.2. Mínimos cuadrados parciales (PLS)

Es una metodología matemática que ha sido diseñada para crear un modelo estadístico el cual permite relacionar variables independientes X(imágenes) con distintas variables dependientes Y(fisicoquímicas) (Sarria, 2018). Este modelo brinda información relevante de los dos conjuntos de datos mediante el desarrollo de un modelo de regresión sobre variables observadas indirectamente (Gómez, 2005). Cuando hay diversos factores y lo que se busca es predecir las variables respuesta, el método va a tener el siguiente modelo estadístico (Sarria, 2018).

$$Y = X\beta + E$$

Donde:

Y= matriz **n*m** la cual va a tener los **n** valores estandarizados de las **m** variables dependientes.

X= matriz **n*p** el cual va a tener los valores estandarizados de la **p** variables predictoras.

β= matriz **p*m** de los parámetros del modelo.

E= matriz **n*m** de errores (Sarria, 2018).

3.3.3. K-vecinos más cercanos (KNN)

Es un método sencillo de clasificación no paramétrico, tiene como fundamento clasificar un registro de datos t, por lo que se recuperan sus k vecinos más cercanos conformando una vecindad de t, cuando se utiliza este método es necesario saber elegir un valor adecuado para k debido a que la clasificación dependerá de este valor. Existen diferentes maneras de seleccionar un valor de k, sin embargo, el más utilizado es mediante la ejecución del algoritmo con diferentes valores de k lo que conlleva a elegir el que presente un mejor rendimiento (Guo et al., 2003). K (número de muestras) puede cambiar de acuerdo a la densidad de puntos, por lo que la distancia que se utilice puede tener cualquier medida; sin embargo, la más utilizada es la euclidiana (Alvídrez, 2019).

3.3.4. KNN Regresión

Es un algoritmo que acumula todos los casos disponibles (anteriores) y los utiliza para predecir los valores de acuerdo a una medida de similitud. Usa la similitud de características para pronosticar lo valores de datos de prueba/nuevos puntos de datos, el valor del nuevo punto se otorga en función de la similitud con otros ejemplos de datos de entrenamiento. La regresión tiene dos enfoques, el primero es mediante el cálculo del promedio del objetivo de los k-vecinos más cercanos, y el segundo es mediante el cálculo ponderado de distancia inversa de los k-vecinos más cercanos. De la misma manera, utiliza las mismas funciones de distancia que la clasificación KNN que son: Euclidiana, Manhattan y Minkowski (Patnaik et al., 2019).

4. CAPÍTULO 4: Parámetros de calidad del modelo

4.1. Coeficiente de determinación (R^2)

Tiene como intervalo de 0-1, los valores altos de R^2 nos indica que la ecuación ajustada representa una buena relación entre Y y X (Bouza, 2018) Este coeficiente es un tipo de medida adimensional, además de que su cálculo e interpretación son fáciles por el intervalo que tiene; sin embargo, el uso común conlleva a interpretaciones abusivas y erróneas; es por eso que es recomendable tenerlo en cuenta como primera medida a completar con otras para que el modelo lineal de regresión ajustado se pueda evaluar y para que las conclusiones que se vayan a realizar sean certeras (Martínez, 2005).

Notar que:

- R^2 es igual al coeficiente de correlación entre las variables.
- 1-R^2 nos indica el porcentaje de variación el cual se explica mediante el modelo de regresión ajustado.
- La mayoría de las veces, el coeficiente se expresa como porcentaje o como 100R^2 (Bouza, 2018).

4.2. R^2 Cross Validado

La Cross-Validación (CV) es una herramienta práctica que sirve para autentificar la predictividad, debido a que valora imparcialmente los errores de predicción; sin embargo, puede llegar a presentar algunos problemas como los modelos de calibración sobreajustados; este tipo de validación se divide en dos subgrupos los cuales constan de: entrenamiento para construir el modelo de regresión y el de predicción. Para todas las muestras se presenta el estadístico "raíz cuadrada del error promedio de CV (RMSECV)" el cual brinda información sobre la capacidad predictiva futura del modelo (Gómez, 2005).

4.3. Coeficiente de correlación cuadrático predictivo (Q^2)

Es uno de los métodos más usados, este coeficiente se utiliza para optimizar un modelo en donde los datos de prueba deben seleccionarse aleatoriamente y deben ser independientes uno de otros. La validación de los modelos a través de datos que no fueron utilizados para el cálculo del modelo, por lo general se denominan como validación externa. Asimismo, mediante el uso de muestras de prueba externas es posible determinar

si un modelo estimado a partir del trainig set puede ser un modelo adecuado para futuros datos candidatos (Consonni et al., 2010).

RESULTADOS Y DISCUSIÓN

Diagrama de Flujo

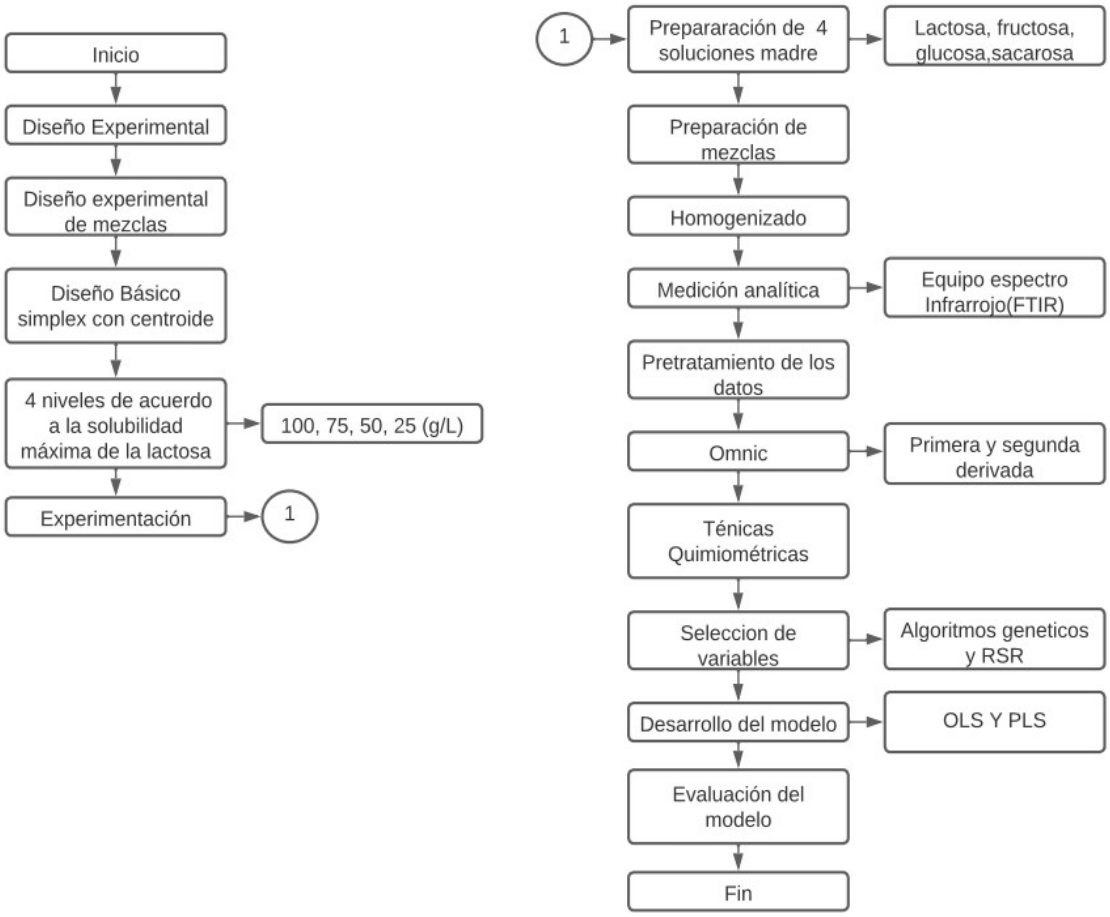


Figura 3. Diagrama de Flujo del diseño experimental de la lactosa.

Condiciones del Diseño Experimental

En este trabajo se utilizó el diseño experimental de mezclas (tabla 2) el cual sirvió de guía para el desarrollo de los patrones para generar los datos y desarrollar los posibles modelos. Se aplicó el diseño básico simplex con centroide para los 4 azúcares los cuales fueron: lactosa, fructosa, glucosa y sacarosa cuya finalidad fue obtener una mezcla equitativa de estos. La experimentación se realizó en base a la solubilidad máxima de la lactosa debido a que presenta una baja solubilidad a comparación con los azúcares mencionados anteriormente (50 a 100 g/l) (National Toxicology Program et al., 1992) de igual manera,

este azúcar comparte la misma región infrarroja lo que conlleva a interferencias en cuanto a su cuantificación (Pérez, 2017).

Se trabajó con 4 niveles del diseño experimental, correspondientes a las siguientes concentraciones 100, 75, 50 y 25 g/L, para cada nivel respectivamente (tabla 1). Los 3 primeros azúcares se disolvieron a temperatura ambiente (20ºC), mientras que la lactosa se disolvió a 50 ºC, una vez realizadas las mezclas acordes al diseño experimental se procedió a realizar las mediciones analíticas.

Tabla 1. Concentraciones de los 4 azúcares correspondiente a cada nivel.

	Concentraciones g/ml			
Nivel	**Lactosa**	**Fructosa**	**Glucosa**	**Sacarosa**
1	50/500	50/500	50/500	50/500
2	37.5/500	37.5/500	37.5/500	37.5/500
3	25/500	25/500	25/500	25/500
4	12.5/500	12.5/500	12.5/500	12.5/500

Tabla 2. Diseño experimental de mezclas.

Lactosa		Fructosa		Glucosa		Sacarosa	
50/500	g/ml	0	ml	0	ml	0	ml
0	ml	50/500	g/ml	0	ml	0	ml
0	ml	0	ml	50/500	g/ml	0	ml
0	ml	0	ml	0	ml	50/500	g/ml
25	ml	25	ml	0	ml	0	ml
25	ml	0	ml	25	ml	0	ml
25	ml	0	ml	0	ml	25	ml
0	ml	25	ml	25	ml	0	ml
0	ml	25	ml	0	ml	25	ml
0	ml	0	ml	25	ml	25	ml
10	ml	10	ml	10	ml	0	ml
10	ml	10	ml	0	ml	10	ml
10	ml	0	ml	10	ml	10	ml
0	ml	10	ml	10	ml	10	ml
25	ml	25	ml	25	ml	25	ml
50	ml	10	ml	10	ml	10	ml
10	ml	50	ml	10	ml	10	ml
10	ml	10	ml	50	ml	10	ml
10	ml	10	ml	10	ml	50	ml
25	ml	25	ml	25	ml	25	ml
25	ml	25	ml	25	ml	25	ml

Mediciones IR

Las mediciones de los espectros se realizaron en el espectrofotómetro Thermo Scientifíc Nicolet Summit PRO FTIR Spectrometer a números de onda de 8000-350 cm^{-1} (región MIR). El espectro fotómetro posee un detector DTGS con una resolución espectral de 0,45 cm^{-1} que le permite identificar distintos componentes en pequeñas cantidades. La técnica utilizada para la recolección de señales fue la técnica de Attenuated Total Reflection (ATR) con ventana de diamante utilizada en el IR.

En esta investigación se obtuvo un total de 84 espectros infrarrojos de las mezclas desarrolladas en el diseño experimental.

Tratamiento de Espectros

El pretratamiento de los datos se ejecutó con el uso del software Omnic (Thermo Electron Corporation, 2006) mediante este programa se obtuvieron datos de la primera y segunda derivada correspondientes a los espectros de las mezclas de los 4 azúcares, los datos que se obtuvieron, sirvieron para posteriormente emplear las técnicas Quimiométricas.

Para esto, se efectuaron las siguientes funciones: Automatic Baseline Correct, Automatic Smooth y Normalice Scale; seguidamente a estos datos guardados se realizó el mismo proceso, pero primero se efectuó la función Display Limits en un rango de 1500-980, consecutivamente a estos nuevos datos generados se eliminó la zona del espectro que tiene influencia del agua lo que conllevó a ejecutar las siguientes funciones View, Display Limits (1500-980), Process, Subtract, Remplace the original spectrum y Save.

Finalmente, se calculó la primera y la segunda derivada de los espectros obtenidos utilizando la función del software Omnic.

Las funciones utilizadas en el software se detallan a continuación:

- Automatic Baseline Correct: Permite corregir la línea de base inclinada de los espectros seleccionados con los puntos de línea de base seleccionado por el software. Esta función realiza un ajuste cuadrático del espectro Y(x) mediante el uso de mínimos cuadrados lineales (Thermo Electron Corporation, 2006).

$$Y(x) \sim Y'(x) = ax^2 + bx + c$$

- Automatic Smooth: Esta función permite perfeccionar la apariencia de los espectros escogidos, mediante la suavización automática del componente que presenta una alta frecuencia cuya finalidad es mejorar la visualización de los picos

oscurecidos por el ruido. Utiliza el filtro de Savistky-Golay con 15 puntos lo que genera un espectro Y', para luego calcular dos factores de escala (SF) (Thermo Electron Corporation, 2006).

$$SF_i = \frac{K}{\frac{\%T * 2.302585}{100}}$$

Donde:

K: Es la ecuación de frecuencia que va a depender del rango en el que se hace la suavización (Thermo Electron Corporation, 2006).

- Normalice Scale: Permite transformar la escala del eje Y de los espectros elegidos a una escala normal en la que los valores de Y de los puntos de datos van desde 0 unidades de absorbancia para el punto más bajo, hasta 1 unidad de absorbancia para el pico más alto (Thermo Electron Corporation, 2006).

  Está representada por la siguiente ecuación:

$$Normalizado = \frac{(Abs - AbsMin)}{(AbsMax - AbsMin)}$$

- Display Limits: Esta opción permite establecer límites a longitudes de onda específicas que sirve para determinar compuestos concretos en base a lo que el analista desee estudiar.
- Primera derivada: Este función permite visualizar los picos que aparecen como bandas u hombros en los espectros originales, es decir, sus bandas se hacen más angostas lo que conlleva a una mejor visualización (Thermo Electron Corporation, 2006).

  Esta derivada se calcula mediante la siguiente ecuación

$$D_i \frac{Y_{i-1} - Y_{i+1}}{X_{i-1} - Y_{i+1}}$$

Donde:

D_i= Primera derivada en cada punto i

Y_i= Absorbancia en cada punto i

X_i= Ubicación en cada punto i (Thermo Electron Corporation, 2006).

- Segunda derivada: Permite encontrar las ubicaciones exactas de los picos en los espectros originales, su cálculo se realiza mediante la siguiente ecuación (Thermo Electron Corporation, 2006).

$$D'_i \frac{Y_{i-2} - 2 * Y_i + Y_{i+2}}{X_{i-1} - X_{i+1}}$$

Donde:

D'_i= Segunda derivada en cada punto i

Y_i= Absorbancia en cada punto i

X_i= Ubicación en cada punto i (Thermo Electron Corporation, 2006).

Partición de Datos

Para el desarrollo del modelo, fue necesario dividir la base de datos en dos partes, la primera parte estaba constituida por el training set el cual sirvió para el desarrollo de la etapa de selección de variables y de modelos, mientras que la segunda parte estaba conformada por el test set el cual permitió validar los modelos provenientes del training set. Para esta partición se realizó una división aleatoria de los datos correspondiendo el 70% de datos para training set y el 30% restante para el test set lo que favorece que la estructura del test set sea similar a la del training set.

Selección de variables

En el presente trabajo, la selección de variables fue efectuada mediante Algoritmos genéticos y el remplazo secuencial renovado (RSR) cuya finalidad fue obtener mayor información mediante la eliminación de variables que contenían información irrelevante además de eliminar el ruido que conducía a la obtención de información falsa en cuanto a la predicción del modelo; este método se realizó con el uso del programa Matlab (Hunt et al., 2001) el cual utiliza varios lenguajes de programación teniendo distintas finalidades de acuerdo a lo que analista desee.

Cálculo del modelo

En el presente trabajo, el modelo se calculó por medio de modelos multivariantes de regresión los cuales fueron: Mínimos cuadrados ordinales (OLS) y Mínimos cuadrados parciales (PLS), estos modelos permitieron tener una idea clara sobre cuan diferentes o semejantes son, pese a tener diferentes variables.

Validación de los modelos

La validación de los modelos se llevó a cabo mediante el coeficiente de determinación R^2 y el coeficiente de correlación cuadrático predictivo (Q^2).

Coeficiente de determinación lineal R^2

Permite dar una idea clara sobre el ajuste de un modelo con la variable que se esté estudiando (López, 2017).

Está representada por la siguiente ecuación

$$R^2 = \frac{SS_{Reg}}{S_{YY}} = 1 - \frac{SSE}{SST}$$

Donde:

SS_{Reg}= Razón de la suma de los cuadrados explicada por la regresión.

S_{YY}= Suma total de cuadrados de desviación sobre la media.

SSE= Suma residual de los cuadrados.

SST= Suma Total de cuadrados (Asuero et al., 2006).

Coeficiente de correlación cuadrático predictivo (Q^2)

Es utilizado como una manera de validar externamente los modelos, este coeficiente es utilizado para la optimización del modelo (Consonni et al., 2010).

Está representada por la siguiente ecuación:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = 1 - \frac{\text{PRESS}}{\text{TSS}}$$

Donde:

n= Número total de objetos en todo el conjunto de datos.

TSS= Suma de las desviaciones al cuadrado de la media del conjunto de datos.

PRESS= Suma de los cuadrados de los errores de predicción (Consonni et al., 2010).

En este trabajo, la base de datos se dividió en 3 partes la cual constó de las siguientes: espectros procesados, primera derivada y segunda derivada.

Mediante el cálculo de cada modelo se escogieron los mejores mediante los valores de calidad del modelo R^2 y Q^2 para posteriormente realizar la fusión y la depuración de los datos con la finalidad de que el modelo mejore. En cuanto a la lactosa, la fusión y la

depuración se realizó mediante la combinación de los 3 modelos obtenidos tanto de los espectros procesados, primera derivada y segunda derivada; mientras que, para el resto de azúcares la fusión se efectuó mediante la combinación de dos mejores modelos provenientes de los espectros originales, primera derivada y segunda derivada, así mismo, la depuración se efectuó a partir de los modelos obtenidos de la fusión.

Los modelos calculados para los métodos de regresión OLS, PLS y RSR para cada azúcar, se muestran en las tablas 4, 5 y 6. En cada tabla se puede observar la selección de variables utilizada siendo algoritmos genéticos GA y el algoritmo de remplazo secuencial renovado RSR, así mismo, se indica el método de regresión dando lugar a los mínimos cuadrados ordinales OLS, mínimos cuadrados parciales PLS, además del método de los k-vecinos más cercanos (KNN-5) para la fructosa.

En la tabla 3, se puede observar las abreviaciones utilizadas en las posteriores tablas.

***Tabla 3*.** Abreviaciones y nomenclatura utilizada en los resultados.

GA	Algoritmos Genéticos
RSR	Algoritmo de Remplazo Secuencial Renovado
OLS	Mínimos Cuadrados Ordinales
PLS	Mínimos Cuadrados Parciales
R^2	Coeficiente de Determinación Lineal
Q^2	Coeficiente de Correlación Cuadrático Predictivo
SN	Sin Resultado
D0	Espectro Procesado
D1	Primera Derivada
D2	Segunda Derivada

Tabla 4. Resultados completos de los espectros procesados correspondientes a cada azúcar.

ESPECTRO PROCESADO													
Azúcar	**Lactosa**			**Fructosa**				**Glucosa**			**Sacarosa**		
Selección de variables	GA		RSR	GA		RSR		GA		RSR	GA		RSR
Método de regresión	OLS	PLS	OLS	OLS	PLS	OLS	KNN-5	OLS	PLS	OLS	OLS	PLS	OLS
R^2	0.534	0.579	0.757	0.609	0.748	0.928	0.849	0.682	0.622	0.730	SR	0.666	0.854
Q^2	0.364	0.548	0.719	0.621	0.697	0.191	0.611	0.365	0.499	0.647	SR	0.558	0.045

Tabla 5. Resultados completos de la primera derivada correspondiente a cada azúcar.

PRIMERA DERIVADA													
Azúcar	**Lactosa**			**Fructosa**				**Glucosa**			**Sacarosa**		
Selección de variables	GA		RSR	GA		RSR		GA		RSR	GA		RSR
Método de regresión	OLS	PLS	OLS	OLS	PLS	OLS	KNN-5	OLS	PLS	OLS	OLS	PLS	OLS
R^2	0.597	0.819	0.841	0.804	0.816	0.899	0.714	0.544	0.458	0.801	0.714	0.866	0.901
Q^2	0.380	0.638	-0.207	0.606	0.517	0.342	0.596	-0.071	-0.227	-0.302	0.699	0.249	0.155

Tabla 6. Resultados completos de la segunda derivada correspondiente a cada azúcar.

SEGUNDA DERIVADA													
Azúcar	**Lactosa**			**Fructosa**				**Glucosa**			**Sacarosa**		
Selección de variables	GA		RSR	GA		RSR		GA		RSR	GA		RSR
Método de regresión	OLS	PLS	OLS	OLS	PLS	OLS	KNN_5	OLS	PLS	OLS	OLS	PLS	OLS
R^2	0.654	0.706	0.875	0.691	0.754	0.833	0.684	0.762	0.681	0.815	0.747	0.771	0.865
Q^2	0.241	0.626	-0.628	0.338	0.639	0.726	0.831	-0.222	-0.147	-0.972	0.413	0.466	0.6

En la tabla 7 se indican los mejores modelos obtenidos a partir de las tablas 4, 5 y 6 teniendo como finalidad la posterior fusión de los datos.

Tabla 7. Mejores modelos OLS, PLS, RSR y KNN-5 para la fusión de los datos.

Azúcar	**Lactosa**									**Fructosa**		**Glucosa**		**Sacarosa**	
Selección de variables	GA						RSR			RSR		GA	RSR	GA	
Método de regresion	OLS			PLS			OLS			OLS	KNN-5	PLS	OLS	PLS	OLS
Derivadas	D0	D1	D2	D0	D1	D2	D0	D1	D2	D2		D0		D0	D1
R^2	0.534	0.597	0.654	0.579	0.819	0.706	0.757	0.841	0.875	0.833	0.684	0.622	0.730	0.666	0.714
Q^2	0.364	0.380	0.241	0.548	0.638	0.626	0.719	-0.207	-0.628	0.726	0.831	0.499	0.647	0.558	0.699

De igual manera, en las tablas 8 y 9 se puede observar la fusión y la depuración de los datos, calculados mediante los distintos métodos de regresión.

Tabla 8. Resultados completos de la fusión, correspondiente a cada azúcar.

FUSIÓN												
Azúcar	**Lactosa**			**Fructosa**			**Glucosa**			**Sacarosa**		
	GA-RSR											
	OLS/PLS (D0, D1, D2)			KNN-5_D2 – RSR_D2			PLS_D0 - RSR_D0			PLS_D0 – OLS_D1		
Selección de variables	GA		RSR	GA		RSR	GA		RSR	GA		RSR
Método de regresión	OLS	PLS	OLS	OLS	PLS	OLS	OLS	PLS	OLS	OLS	PLS	OLS
R^2	0.655	0.783	0.833	0.908	0.905	0.904	0.506	0.598	0.694	0.678	**0.749**	0.812
Q^2	0.549	0.649	-0.151	0.805	0.844	0.802	0.492	0.485	0.7	0.746	**0.795**	0.522

Tabla 9. Resultados completos de la depuración, correspondiente a cada azúcar.

DEPURACIÓN													
Azúcar	**Lactosa**				**Fructosa**			**Glucosa**			**Sacarosa**		
Selección de variables	GA		RSR		GA		RSR	GA		RSR	GA		RSR
Método de regresión	OLS	PLS	OLS	KNN-5	OLS	PLS	OLS	OLS	PLS	OLS	OLS	PLS	OLS
R^2	0.76	**0.773**	* [1]	**0.802**	**0.865**	0.868	SR	SR	**0.771**	0.731	0.703	0.81	0.869
Q^2	0.633	**0.726**	*	**0.731**	**0.824**	0.783	SR	SR	**0.81**	0.093	0.53	0.67	0.6

En la tabla 10, se visualiza los mejores modelos que se obtuvieron mediante la fusión y la depuración de los datos, validados mediante los valores de calidad del modelo R^2 y Q^2.

Tabla 10. Mejores modelos obtenidos para cada azúcar después de la fusión y la depuración de los datos.

Azúcar	**Lactosa**		**Fructosa**	**Glucosa**	**Sacarosa**
	Depuración				Fusión
Selección de variables	GA	RSR	GA	GA	GA
Método de regresión	PLS	KNN-5	OLS	PLS	PLS
R^2	0.773	0.802	0.865	0.771	0.749
Q^2	0.726	0.731	0.824	0.81	0.795

[1] Los datos en predicción de la fusión fueron malos por lo que no se tomaron en cuenta para la depuración.

Tabla 11. Variables seleccionadas en cada uno de los mejores modelos.

Azúcar	**Lactosa**			**Fructosa**	**Glucosa**	**Sacarosa**	
Selección de variables	GA / RSR			GA	GA	GA	
Método de regresión	PLS / KNN-5			OLS	PLS	PLS	
Derivadas	D0	D1	D2	D2	D0	D0	D1
Variables (λ)	1093.923, 1020.641, 1023.052, 1091.994	1096.333, 1097.298, 1096.816, 1176.365	1106.94, 1107.422, 1107.904, 1242.897	1066.924, 1063.067, 1065.478, 1086.209, 992.678, 1076.567	1166.722, 999.910, 1011.481, 1167.205, 982.554, 1003.767, 1009.07, 1115.618	1131.046, 1057.282, 1132.492	984.0002, 988.8214, 1274.717

CONCLUSIONES

En el trabajo realizado se puede evidenciar que se encontraron dos mejores modelos para la lactosa con el uso de algoritmos genéticos para la selección de variables y PLS como método de regresión, teniendo como valores de calidad del modelo R^2= 0.773 y Q^2= 0.726 y el segundo utilizando Algoritmo de Remplazo Secuencial Renovado como método de selección de variables y con el algoritmo KNN-5 como método de regresión presentando valores de calidad del modelo R^2= 0.802 y Q^2= 0.731.

Al realizar una comparación de los resultados con la tesis realizada por Pérez (Pérez, 2017) se pudo observar que los modelos para la fructosa, glucosa y sacarosa fueron mejores a comparación de los modelos obtenidos en esta tesis, debido a que en el trabajo anterior se centró de manera general en todos estos azúcares; conllevando a no obtener un modelo para la lactosa; esto fue debido a que los azúcares comparten la misma región infrarroja por lo que tienden a provocar interferencias (figura 2). Es por eso que este trabajo fue enfocado principalmente en la solubilidad de la lactosa lo que ocasionó complicaciones al momento del cálculo de los modelos.

En cuanto a la Fructosa, el mejor modelo fue mediante algoritmos genéticos para la selección de variables y OLS como método de regresión presentando valores de calidad del modelo R^2= 0.865 y Q^2= 0.824. De igual manera, el mejor modelo para la Glucosa fue mediante los algoritmos genéticos para la selección de variables y PLS como método de regresión presentando valores de calidad del modelo R^2= 0.771 y Q^2= 0.81. Mientras que, para la Sacarosa, el mejor modelo que se obtuvo fue al realizar la fusión de los datos con la selección de variables que fue el algoritmo genético por el método PLS como método de regresión, con valores de calidad del modelo de R^2= 0.749 y Q^2= 0.795.

Así mismo, se puede concluir que si vale la pena realizar la fusión y la depuración de los datos porque el modelo si mejora. En el caso de la fusión, se combinaron las variables de los mejores modelos lo que en ciertos casos puede conllevar a obtener mejores resultados, tal fue el caso de la sacarosa pudiendo evidenciar que el mejor modelo para este azúcar fue mediante la fusión de los datos presentados en la tabla 10 ; de igual manera, en la misma tabla se evidencia que los mejores modelos obtenidos para la lactosa, fructosa y glucosa fue mediante la depuración de los datos la cual consistió en eliminar los datos que se encontraban lejos del espectro global por lo que los resultados fueron mejores. Cabe indicar que, el hecho de tener que haber utilizado la primera y segunda derivada

para obtener un mejor modelo y fusionar estos datos, indica que no toda la información necesaria para el desarrollo de un modelo para la lactosa se la puede obtener el espectro original. Mucha de la información se encuentra oculta por la interacción de los otros azúcares y porque estos comparten la misma zona del espectro y al derivar el espectro esta se vuelve evidente, por esta razón el modelo de la lactosa las utiliza.

Finalmente, en lo que respecta a las longitudes de onda, en la tabla 11 se puede observar que, para la lactosa, los dos mejores modelos correspondieron a las variables 1093.923, 1020.641, 1023.052, 1091.994 correspondientes a los datos originales, 1096.333, 1097.298, 1096.816, 1176.365 correspondientes a la primera derivada y las variables 1106.94, 1107.422, 1107.904, 1242.897 correspondientes a la segunda derivada. En la misma tabla se puede observar las longitudes de onda para los mejores modelos de la Fructosa, Glucosa y Sacarosa.

REFERENCIAS BIBLIOGRÁFICAS

Abraham, K., Schlumbach, K., Rohde, A., & Flöter, E. (2014). *Análisis y Ventajas de la Hidrólisis Enzimática del Dextrano en el Proceso del Azúcar*. https://atamexico.com.mx/wp-content/uploads/2017/11/5-ELABORACI%C3%93N-2016.pdf

Alvídrez, F. (2019). *Biosensado para la detección de emociones: Clasificación de eventos para dos tecnologías.* [[Universidad Autónoma de Chihuahua]]. http://repositorio.uach.mx/247/1/BIOSENSADO%20PARA%20LA%20DETECCI%C3%93N%20DE%20EMOCIONES%20-%20CLASIFICACI%C3%93N%20DE%20EVENTOS%20PARA%20DOS%20TECNOLOG%C3%8DAS%20_%20FLORENTINO%20ALV%C3%8DDREZ.pdf

Asuero, A. G., Sayago, A., & González, A. G. (2006). The correlation coefficient: An overview. In *Critical Reviews in Analytical Chemistry* (Vol. 36, Issue 1). https://doi.org/10.1080/10408340500526766

Badui, S. (2013). *Química de los alimentos* (5ta ed.). Pearson.

Bayas, A., & Saltos, H. (2010). *Aplicación de un Diseño Experimental de Mezclas en el Desarrollo de una "'Barra Energética'" con base en el Salvado de Palmito de Pejibaye (Bactris gasipaes H.B.K)*. Revista Tecnológica ESPOL-RTE. http://200.10.150.204/index.php/tecnologica/article/viewFile/48/20

Billabio, D., & Todeschini, R. (2009). *Multivariate Classification for Qualitative Analysis.* (D.-W. Sun, Ed.). Infrared Spectroscopy for Food Quality Analysis and Control. https://doi.org/10.1016/B978-0-12-374136-3.00004-3

Bouza, C. (2018). *Modelos de Regresión y sus Aplicaciones*.

Camacho, A. (2018). *Análisis de alta tecnología para azúcares reductores*.

Camino, M., Miguez, G., Sánchez, K., & Zurita, J. (2020). *Método de Munson y Walker* [Universidad Técnica de Ambato]. https://www.studocu.com/ec/document/universidad-ute/quimica-analitica/ape-14-na-metodo-de-munson-y-walker/9694779

Cassotti, M., Grisoni, F., & Todeschini, R. (2014). Reshaped Sequential Replacement algorithm: An efficient approach to variable selection. *Elsevier*, 137.

Consonni, V., Ballabio, D., & Todeschini, R. (2010). Evaluation of model predictive ability by external validation techniques. *Journal of Chemometrics*. https://doi.org/10.1002/cem.1290

Contreras, E., Carvajal, M., & Castro, M. (2019). *Determinación de azúcares reductores y totales por el método de Lane-Eynon*. SCRIB.

Criollo, T., & Cueva, M. (2017). *Cuantificación de Glucosa, Fructosa y Sacarosa en muestras de pulpa y jugo de uvilla (Physallis peruviana), Tomate de árbol (Solanum betaceum) y Manzana (Pyrus malus) por Espectroscopia Infrarroja por Transformadas de Fourier (FTIR)* [Universidad de Cuenca].

https://dspace.ucuenca.edu.ec/bitstream/123456789/27129/1/Trabajo%20de%20Titulaci%C3%B3n.pdf

Cuevas, E., Fausto, F., Gálvez, J., & Rodríguez, A. (2021). *Matlab: computación meteheurística y bio-inspirada*. RC Libros.

de Fuentes Navarta, M., Bosch Ojeda, C., & Sánchez Rojas, F. (2008). *Aplicación de la Espectroscopia del Infrarrojo Medio en Química Analítica de Procesos.* Boletín de la Sociedad Química de México, 2(3). http://bsqm.org.mx//pdf-boletines/V2/N3/1-Fuentes Navarta.pdf

Delgado, J. (2008). *Optimización de la formulación y elaboración de salsa de tomate picante mediante diseño experimental de mezclas y variables de procesos* [Universidad del Azuay]. http://dspace.uazuay.edu.ec/handle/datos/7682

Domínguez, A. (2014). *Clasificación Automatizada de Rones Havana Club y Santero mediante la Espectrofotometría Ultravioleta-Visible y la Quimiometría* [Universidad Central Marta Abreu de las Villas]. https://dspace.uclv.edu.cu/bitstream/handle/123456789/110/Andy%20G.%20Dom%C3%ADnguez%20Rodr%C3%ADguez.pdf?sequence=1&isAllowed=y

Flores, O. (2017). *Estudio de la Composición Nutricional de un Alimento por Espectroscopía de Infrarrojo cercano con Transformada de Fourier (FT-NIRS) y Quimiometría* [Universidad Autónoma de Baja California]. https://repositorioinstitucional.uabc.mx/bitstream/20.500.12930/2551/1/MXL120651.pdf

Gómez, M. (2005). *Aplicaciones al control de calidad industrial de la espectroscopia infrarroja media combinada con métodos quimiométricos multivariantes*.

Guo, G., Wang, H., Bell, D. A., Bi, Y., Bell, D., & Greer, K. (2003). *KNN Model-Based Approach in Classification* . https://doi.org/10.1007/978-3-540-39964-3_62

Gutiérrez, H., & de la Vara, R. (2008). *Análisis y diseño de experimentos* (2da ed.). Mc Graw Hill.

Hernández, A., Sánchez, D., Sánchez, F., Zuñiga, Z., García, I., Dinkova, T., & Avila, A. (2020). *Quantification of Reducing Sugars Based on the Qualitative Technique of Benedict*. https://doi.org/10.1021/acsomega.0c04467

Herrera, A. (2011). *Estudio comparativo de métodos para la determinación de sacarosa y azúcares reductores en miel virgen de caña utilizados en el ingenio Pichichí S.A.* [Universidad Tecnológica de Pereira]. https://repositorio.utp.edu.co/server/api/core/bitstreams/4096319e-f5c0-4f31-a4b6-4d16bc040bff/content

Hunt, B., Lipsman, R., Rosenberg, J., Coobes, K., Osborn, J., & Stuck, G. (2001). *A guide to Matlab for beginners and experienced users*. http://www.uop.edu.pk/ocontents/A%20Guide%20to%20MATALB.pdf

Kuaquira, S., & Huaman, L. (2022). *Diseño e implementación de un sistema de identificación de 5 tipos de material plástico mediante espectroscopía de*

reflectancia infrarroja cercana (NIR) [Universidad Nacional de San Antonio Abad del Cusco]. https://repositorio.unsaac.edu.pe/bitstream/handle/20.500.12918/6400/253T20220073_TC.pdf?sequence=1

López, J. (2017). *Coeficiente de determinación (R cuadrado)*. Economipedia. https://economipedia.com/definiciones/r-cuadrado-coeficiente-determinacion.html

Martínez, E. (2005). Errores frecuentes en la interpretación del coeficiente de determinación lineal. *Anuario Jurídico y Económico Escurialense*.

Massart, D., Vandeginste, B., Buydens, L., de Jong, S., Lewi, P., & Smeyers, J. (1998). *Handbook of Chemometrics and Qualimetrics: Part A*. Elseiver.

Maurad, L. (2016). *Identificación de ácidos grasos trans presentes en galletas industriales y artesanales mediante espectrometría infrarroja* [Universidad del Azuay]. https://dspace.uazuay.edu.ec/bitstream/datos/5730/1/12050.pdf

Mondragón, P. (2017). *Espectroscopía de Infrarrojo para Todos... y 51 Espectros de Alimentos Consumidos en México*. Ciatej.

National Toxicology Program, Institute of Environmental Health Sciences, & National Institutes of Health (NTTP). (1992). *Compound Summary beta-lactose*. https://pubchem.ncbi.nlm.nih.gov/compound/6134#section=Solubility

Patnaik, S., Yang, X., & Sethi, I. (2019). *Advances in Machine Learning and Computational Intelligence. Procedings of IMLCI.*

Pérez, A. (2017). *Tratamiento Matemático de señales Espectrofotométricas Infrarrojas para la Cuantificación de Azúcares* [Universidad del Azuay]. https://dspace.uazuay.edu.ec/bitstream/datos/6854/1/12825.pdf

Reyes, M. (2009). *Aplicación del diseño experimental en el desarrollo de las prácticas internas, en el área de operaciones unitarias.* [Universidad de San Carlos de Guatemala]. http://biblioteca.usac.edu.gt/tesis/08/08_1138_Q.pdf

Rodriguez, L., & Allendorf, M. (2011). Use of FTIR for rapid authentication and detection of adulteration of food. *Review of Food Science and Tecnology*. https://doi.org/10.1146/annurev-food-022510-133750

Rojas, E., & Rojas, L. (2000). *Exploración al Diseño Experimental.* Neogranadina.

Sarria, J. (2018). *Aplicación de la técnica de análisis de imágenes en el espectro visible para la optimización y control del tostado de café.* [Universidad Surcolombiana]. http://repositoriousco.co:8080/jspui/handle/123456789/1021

Serrano, J. (2017). *Espectroscopía Infrarroja 1-Fundamentos.* https://www.upct.es/~minaeees/espectroscopia_infrarroja.pdf

Téllez, C. (2019). *Aplicaciones de la Espectroscopía Infrarroja en el Análisis de Alimentos* [Universidad de Sevilla]. https://idus.us.es/bitstream/handle/11441/91690/T%C3%89LLEZ%20MESA%2C%20CLARA.pdf?sequence=1&isAllowed=y

Thermo Electron Corporation. (2006). OMNIC User's Guide. In *Madison*.

Torres, G. (2016). *Estudio del efecto del tipo de poliestireno usado como envase plástico para alimentos sobre la migración global mediante espectroscopia IR-ATR y PCA* [Universidad Nacional de Colombia]. https://repositorio.unal.edu.co/bitstream/handle/unal/58345/ginaalexandratorreslopez.2016.pdf?sequence=1&isAllowed=y

Úbeda, A. (2012). *Análisis del perfil de azúcares en la autentificación de zumos de frutas* [Universidad Politécnica de Cartagena]. https://repositorio.upct.es/bitstream/handle/10317/3143/pfc5012.pdf?sequence=1

Valcárcel, M. (2009). *Optimización del proceso de evaluación y selección de germoplasma de tomate por características de calidad organoléptica: uso de la tecnología NIR y sensores electrónicos*. [Escola Superior de Tecnologia y Ciències Experimentals].

van de Voort, F. R. (1992). Fourier transform infrared spectroscopy applied to food analysis. *Food Research International*, *25*(5). https://doi.org/10.1016/0963-9969(92)90115-L

Zumbado, H. (2010). *Métodos Cromatográficos* [Universidad de la Habana]. https://www.academia.edu/32304048/M%C3%A9todos_Cromatogr%C3%A1ficos